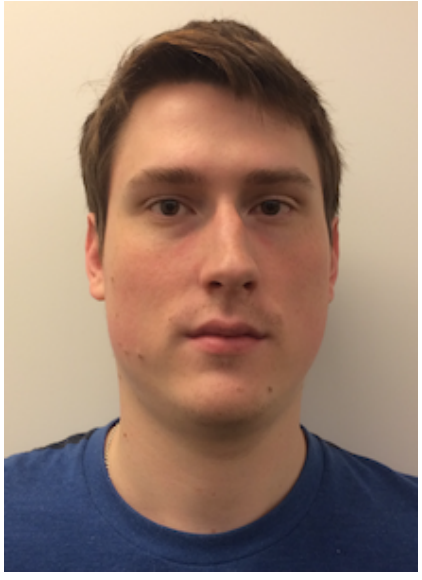
On algorithmic efficiency and statistical optimality in empirical risk minimization

Dmitrii M. Ostrovskii
http://ostrodmit.github.io

TTI-Chicago
April 8, 2019

Part I. Efficient Algorithms for Multiclass Classification

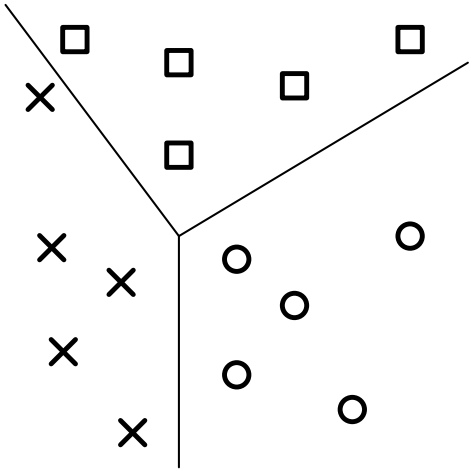
Joint with **Dmitry Babichev** (equal contribution) and **Francis Bach**



D. Babichev, D. M. Ostrovskii, F. Bach. Efficient Primal-Dual Algorithms for Large-Scale Multiclass Classification. *arXiv:1902.03755*.

Multiclass linear classification

- Dataset (x_i, y_i), $i \in [n] = \{1, 2, ..., n\}$.
- $x_i \in \mathbb{R}^d$ features, $y_i \in \{e_1, ..., e_k\}$ class – a vertex of simplex Δ_k.
- **Large-scale:** d, n, k up to $10^6 - 10^9$, as in NLP applications.



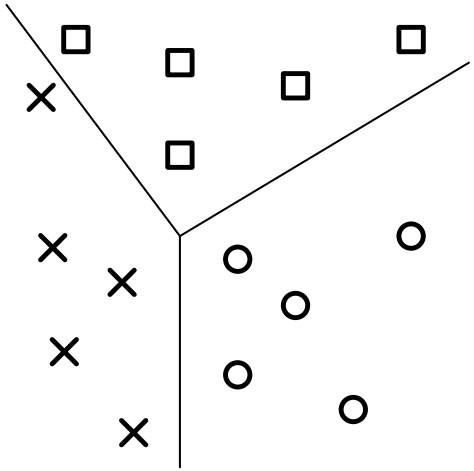
Regularized empirical risk minimization (ERM)

$$\min_{U \in \mathbb{R}^{d \times k}} \frac{1}{n} \sum_{i=1}^{n} \ell(U^\top x_i, y_i) + \underbrace{\lambda \|U\|_{\mathscr{U}}}_{\text{Regularization}} .$$

- **Goal:** $U^* \in \mathbb{R}^{d \times k}$ encoding k classifiers (**empirical minimizer!**)
- **Loss** $\ell : \mathbb{R}^k \times \Delta_k \to \mathbb{R}$ – misfit of predicting the (soft) label $y \in \Delta_k$ from the output of k classifiers $U^\top x \in \mathbb{R}^k$.

Multiclass linear classification

- Dataset (x_i, y_i), $i \in [n] = \{1, 2, ..., n\}$.
- $x_i \in \mathbb{R}^d$ features, $y_i \in \{e_1, ..., e_k\}$ class – a vertex of simplex Δ_k.
- **Large-scale:** d, n, k up to $10^6 - 10^9$, as in NLP applications.



Regularized empirical risk minimization (ERM)

$$\min_{U \in \mathbb{R}^{d \times k}} \frac{1}{n} \sum_{i=1}^{n} \ell(U^\top x_i, y_i) + \underbrace{\lambda \|U\|_{\mathscr{U}}}_{\text{Regularization}} .$$

- **Goal:** $U^* \in \mathbb{R}^{d \times k}$ encoding k classifiers (**empirical minimizer!**)
- **Loss** $\ell : \mathbb{R}^k \times \Delta_k \to \mathbb{R}$ – misfit of predicting the (soft) label $y \in \Delta_k$ from the output of k classifiers $U^\top x \in \mathbb{R}^k$.

Goal: find an optimum U^* with *sublinear* iteration cost $\widetilde{O}(d + n + k)$.

Challenge and previous work

Input: $X \in \mathbb{R}^{n \times d}$, $Y \in \Delta_k^{\otimes n}$; **output**: $U \in \mathbb{R}^{dk} \Rightarrow O(dn + dk + nk)$.

Biclass case: $k = O(1)$, linear time $O(dn)$, sublinear time $O(d + n)$.

- **Dual sampling**: sample $i \in [n]$, compute $\nabla_u \ell(u^\top x_i, y_i)$ in $O(d)$.
  - SGD, Pegasos [Shalev-Shwartz et al., 2011], SVRG [Johnson and Zhang, 2013], SAGA [Defazio et al., 2014], SAG [Schmidt et al., 2017], SDCA [Shalev-Shwartz and Zhang, 2013], Frank-Wolfe [Lacoste-Julien et al., 2012],...

Challenge and previous work

Input: $X \in \mathbb{R}^{n \times d}$, $Y \in \Delta_k^{\otimes n}$; **output**: $U \in \mathbb{R}^{dk} \Rightarrow O(dn + dk + nk)$.

Biclass case: $k = O(1)$, linear time $O(dn)$, sublinear time $O(d + n)$.

- **Dual sampling**: sample $i \in [n]$, compute $\nabla_u \ell(u^\top x_i, y_i)$ in $O(d)$.
  - SGD, Pegasos [Shalev-Shwartz et al., 2011], SVRG [Johnson and Zhang, 2013], SAGA [Defazio et al., 2014], SAG [Schmidt et al., 2017], SDCA [Shalev-Shwartz and Zhang, 2013], Frank-Wolfe [Lacoste-Julien et al., 2012],...
- **Saddle-point approach:** recast as a quasi-bilinear saddle-point problem with variables in $\mathbb{R}^d$ and $\mathbb{R}^n$, solve by mirror descent or mirror prox. Partial gradients in $O(dn)$, accelerated to $O(d + n)$.
  - [Grigoriadis and Khachiyan, 1995; Juditsky and Nemirovski, 2011; Xiao et al., 2017],...
  - Explicit bounds on sampling variance, certificate on duality gap.

Challenge and previous work

Input: $X \in \mathbb{R}^{n \times d}$, $Y \in \Delta_k^{\otimes n}$; **output**: $U \in \mathbb{R}^{dk} \Rightarrow O(dn + dk + nk)$.

Biclass case: $k = O(1)$, linear time $O(dn)$, sublinear time $O(d + n)$.

- **Dual sampling**: sample $i \in [n]$, compute $\nabla_u \ell(u^\top x_i, y_i)$ in $O(d)$.
  - SGD, Pegasos [Shalev-Shwartz et al., 2011], SVRG [Johnson and Zhang, 2013], SAGA [Defazio et al., 2014], SAG [Schmidt et al., 2017], SDCA [Shalev-Shwartz and Zhang, 2013], Frank-Wolfe [Lacoste-Julien et al., 2012],...
- **Saddle-point approach:** recast as a quasi-bilinear saddle-point problem with variables in $\mathbb{R}^d$ and $\mathbb{R}^n$, solve by mirror descent or mirror prox. Partial gradients in $O(dn)$, accelerated to $O(d + n)$.
  - [Grigoriadis and Khachiyan, 1995; Juditsky and Nemirovski, 2011; Xiao et al., 2017],...
  - Explicit bounds on sampling variance, certificate on duality gap.

Multiclass case: all those become $O(dk)$ or $O(dk + nk)$, i.e. *linear*.

Reduction to saddle-point problem

We consider the class of **Fenchel-Young losses** [Blondel et al., 2018]:

$$\ell(U^\top x, y) = \max_{v \in \Delta_k} \Big\{ \underbrace{f_y(v)}_{\text{concave and "simple"}} + \underbrace{(v-y)^\top U^\top x}_{\text{bilinear}} \Big\}.$$

- k-softmax loss (logistic):

$$f_y(v) = h(v) := -\sum_{i \in [n]} v_i \log(v_i)$$

- k-hinge loss (SVM) [Shalev-Shwartz and Ben-David, 2014]:

$$f_y(v) = 1 - v^\top y.$$

Reduction to saddle-point problem

We consider the class of **Fenchel-Young losses** [Blondel et al., 2018]:

$$\ell(U^\top x, y) = \max_{v \in \Delta_k} \Big\{ \underbrace{f_y(v)}_{\text{concave and "simple"}} + \underbrace{(v - y)^\top U^\top x}_{\text{bilinear}} \Big\}.$$

- k-softmax loss (logistic):

$$f_y(v) = h(v) := -\sum_{i \in [n]} v_i \log(v_i)$$

- k-hinge loss (SVM) [Shalev-Shwartz and Ben-David, 2014]:

$$f_y(v) = 1 - v^\top y.$$

Saddle-Point Formulation

$$\min_{\|U\|_{\mathscr{U}} \leqslant R^*} \max_{V \in \mathcal{V}} \underbrace{\frac{1}{n} \mathrm{tr}\left[{}_k(V - Y)^\top {}_n X_d U_k\right]}_{\textbf{bi-affine}} + \underbrace{\lambda \|U\|_{\mathscr{U}}}_{\text{simple}} + \underbrace{\mathcal{F}(V)}_{= \frac{1}{n}\sum_{i \in [n]} f_{y_i}(v_i), \text{ simple}}$$

and the dual feasible set $\mathcal{V} = \Delta_k^{\otimes n}$ is the *direct product of simplices*.

Mirror descent: primer

$$\min_{U \in \mathcal{U}} \underbrace{f(U)}_{\text{cvx, smooth}} + \underbrace{\Psi(U)}_{\text{cvx, simple}}$$

(Composite) **gradient descent:** given stepsizes $\{\gamma_t\}$, iterate

$$U^{t+1} = \operatorname*{argmin}_{U \in \mathcal{U}} \left\{ \langle \nabla f(U^t), U \rangle + \Psi(U) + \frac{1}{\gamma_t} \frac{\|U - U^t\|_2^2}{2} \right\}$$

Mirror descent: replace $\|U - U^t\|_2^2$ with **Bregman divergence**:

$$D_{\phi_{\mathcal{U}}}(U, U^t) := \phi_{\mathcal{U}}(U) - \phi_{\mathcal{U}}(U^t) - \langle \nabla \phi_{\mathcal{U}}(U^t), U - U^t \rangle$$

for some potential $\phi_{\mathcal{U}}(\cdot)$ generalizing $\frac{1}{2}\|\cdot\|_2^2$ and such that:

Mirror descent: primer

$$\min_{U \in \mathcal{U}} \underbrace{f(U)}_{\text{cvx, smooth}} + \underbrace{\Psi(U)}_{\text{cvx, simple}}$$

(Composite) **gradient descent:** given stepsizes $\{\gamma_t\}$, iterate

$$U^{t+1} = \operatorname*{argmin}_{U \in \mathcal{U}} \left\{ \langle \nabla f(U^t), U \rangle + \Psi(U) + \frac{1}{\gamma_t} \frac{\|U - U^t\|_2^2}{2} \right\}$$

Mirror descent: replace $\|U - U^t\|_2^2$ with **Bregman divergence**:

$$D_{\phi_{\mathcal{U}}}(U, U^t) := \phi_{\mathcal{U}}(U) - \phi_{\mathcal{U}}(U^t) - \langle \nabla \phi_{\mathcal{U}}(U^t), U - U^t \rangle$$

for some potential $\phi_{\mathcal{U}}(\cdot)$ generalizing $\frac{1}{2}\|\cdot\|_2^2$ and such that:

- $\phi_{\mathcal{U}}$ is 1-strongly convex w.r.t. the given norm $\|\cdot\|_{\mathcal{U}}$.
- "Fitting" geometry $\Omega_{\mathcal{U}} = \widetilde{O}(\text{radius}^2_{\|\cdot\|_{\mathcal{U}}}(\mathcal{U}))$.
- Step easily computable – usually a quasi-separable problem.

Mirror descent for saddle-point problems

Given convex sets $\mathcal{U}$ and $\mathcal{V}$, consider a saddle-point problem

$$\min_{U\in\mathcal{U}} \max_{V\in\mathcal{V}} \underbrace{f(U,V)}_{\text{smooth, cvx in } U\text{, ccv in } V} + \underbrace{\Psi(U)}_{\text{cvx, simple}} + \underbrace{\mathcal{F}(V)}_{\text{ccv, simple}}$$

- primal-dual variable: $W = (U, V) \in \mathcal{W} = \mathcal{U} \times \mathcal{V}$
- gradient field $G(W) = [\nabla_U f(U,V), -\nabla_V f(U,V)]$
- "balanced" joint potential $\phi_{\mathcal{W}}(W) = \frac{1}{\Omega_{\mathcal{U}}}\phi_{\mathcal{U}}(U) + \frac{1}{\Omega_{\mathcal{V}}}\phi_{\mathcal{V}}(V)$.

Mirror descent

$$W^{t+1} = \mathrm{prox}_{W^t}(G(W^t))$$
$$:= \operatorname*{argmin}_{W\in\mathcal{W}} \left\{ \langle G(W^t), W\rangle + \Psi(U) - \mathcal{F}(V) + \frac{D_{\phi_{\mathcal{W}}}(W, W^t)}{\gamma_t} \right\},$$

Mirror prox [Nemirovski and Yudin, 1983], faster convergence rate:

$$W^{t+1/2} = \mathrm{prox}_{W^t}(G(W^t)),$$
$$W^{t+1} = \mathrm{prox}_{W^t}(G(W^{t+1/2})).$$

Mirror descent for saddle-point problems

Given convex sets $\mathcal{U}$ and $\mathcal{V}$, consider a saddle-point problem

$$\min_{U \in \mathcal{U}} \max_{V \in \mathcal{V}} \underbrace{f(U,V)}_{\text{smooth, cvx in } U\text{, ccv in } V} + \underbrace{\Psi(U)}_{\text{cvx, simple}} + \underbrace{\mathcal{F}(V)}_{\text{ccv, simple}}$$

- primal-dual variable: $W = (U, V) \in \mathcal{W} = \mathcal{U} \times \mathcal{V}$
- gradient field $G(W) = [\nabla_U f(U,V), -\nabla_V f(U,V)]$
- "balanced" joint potential $\phi_{\mathcal{W}}(W) = \frac{1}{\Omega_{\mathcal{U}}}\phi_{\mathcal{U}}(U) + \frac{1}{\Omega_{\mathcal{V}}}\phi_{\mathcal{V}}(V)$.

Mirror descent

$$W^{t+1} = \mathrm{prox}_{W^t}(G(W^t))$$
$$:= \operatorname*{argmin}_{W \in \mathcal{W}} \left\{ \langle G(W^t), W \rangle + \Psi(U) - \mathcal{F}(V) + \frac{D_{\phi_{\mathcal{W}}}(W, W^t)}{\gamma_t} \right\},$$

Stochastic mirror descent: replace $G(W^t)$ with cheaper **stochastic oracle** $\zeta(W^t)$ such that $\mathbf{E}[\zeta(W^t)] = G(W^t)$.

General convergence guarantee

Consider a saddle-point problem with *quasi-bilinear* objective:

$$f(U, V) = \frac{1}{n}\mathrm{tr}\left[(V - Y)^\top XU\right],$$

$$G(W) = \frac{1}{n}[X^\top(V - Y), -XU],$$

$$\zeta(W) = \frac{1}{n}[\eta_{V,Y}; -\xi_U]$$

Given norms $\|\cdot\|_{\mathcal{U}}, \|\cdot\|_{\mathcal{V}}$ and their dual norms $\|\cdot\|_{\mathcal{U}^*}, \|\cdot\|_{\mathcal{V}^*}$, define:

- Cross-Lipschitz constant $\mathcal{L}_{\mathcal{U},\mathcal{V}}$ of $G(\cdot)$:

$$\mathcal{L}_{\mathcal{U},\mathcal{V}} := \frac{1}{n}\sup_{\|U\|_{\mathcal{U}} \leqslant 1} \|XU\|_{\mathcal{V}^*}.$$

- Uniform bounds on "variances" of ξ_U and $\eta_{V,Y}$ over $\mathcal{W}$:

$$\bar{\sigma}^2_{\mathcal{U}} \geqslant \frac{1}{n^2}\sup_{U \in \mathcal{U}} \mathbf{E}[\|XU - \xi_U\|^2_{\mathcal{V}^*}], \quad \bar{\sigma}^2_{\mathcal{V}} \geqslant \ldots$$

General convergence guarantee

Theorem (see [Juditsky and Nemirovski, 2011])

After T iterations of stochastic mirror descent with appropriate constant stepsize and uniform averaging, the expected duality gap is

$$\mathbf{E}[\mathtt{Gap}^T] \lesssim \frac{\mathcal{L}_{\mathcal{U},\mathcal{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}}{\sqrt{T}}$$

- Cross-Lipschitz constant $\mathcal{L}_{\mathcal{U},\mathcal{V}}$ of $G(\cdot)$:

$$\mathcal{L}_{\mathcal{U},\mathcal{V}} := \frac{1}{n} \sup_{\|U\|_{\mathcal{U}} \leqslant 1} \|XU\|_{\mathcal{V}^*}.$$

- Uniform bounds on "variances" of ξ_U and $\eta_{V,Y}$ over $\mathcal{W}$:

$$\bar{\sigma}_{\mathcal{U}}^2 \geqslant \frac{1}{n^2} \sup_{U \in \mathcal{U}} \mathbf{E}[\|XU - \xi_U\|_{\mathcal{V}^*}^2], \quad \bar{\sigma}_{\mathcal{V}}^2 \geqslant \ldots$$

Choice of geometry

Saddle-Point Formulation

$$\min_{U\in\mathcal{U}} \max_{V\in\mathcal{V}} \frac{1}{n}\mathrm{tr}\left[(V-Y)^\top XU\right] + \underbrace{\lambda\|U\|_{\mathscr{U}}}_{\text{simple}} + \underbrace{\mathcal{F}_Y(V)}_{\text{simple}}$$

where $\mathcal{U}$ is an $\|\cdot\|_{\mathscr{U}}$-ball, and $\mathcal{V} = \Delta_k^{\otimes n}$ is the product of simplices.

Now we have to choose the setup $\|\cdot\|_{\mathscr{U}}, \|\cdot\|_{\mathscr{V}}, \phi_{\mathcal{U}}(\cdot), \phi_{\mathcal{V}}(\cdot)$ such that

$$\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}} \approx \sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2},$$

and both are small (hopefully not growing with d, n, k).

Choice of geometry: dual

Mixed norms with $p, q \in [1, \infty]$:

$$\|V\|_{\mathscr{V}} = \|V\|_{p\times q} = \left(\sum_{i\in[n]} \|V(i,:)\|_q^p \right)^{1/p},$$

We would like to choose p, q "compatible" with $\mathscr{V} = \Delta_k^{\otimes n}$. **Intuition:**

- When $n = 1$, we have $\mathscr{V} = \Delta_k$, and $\|\cdot\|_{p\times q} \equiv \|\cdot\|_q \Rightarrow \boxed{q = 1}$.
- When $k = 2$, we have $\mathscr{V} = [-1, 1]^n$ – cube, and $\|\cdot\|_{p\times q} \equiv \|\cdot\|_p$. Correct choice known from [Nemirovski and Yudin, 1983]: $\boxed{p = 2}$.

$$\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}, \quad \phi_{\mathcal{V}}(V) = -\sum_{i=1}^{n} h(V(i,:)),$$

where $h : \Delta_k \to \mathbb{R}$ is Shannon's entropy over rows.

Choice of geometry: primal

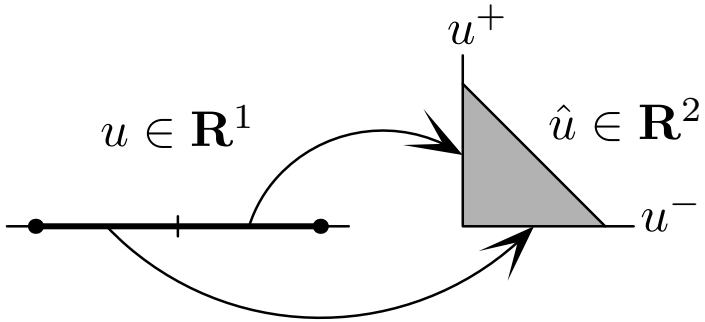
Apart from statistical considerations, we have algorithmic constraints:

- Need $p = 1$ to control variance when sampling features in the biclass case [Juditsky and Nemirovski, 2011].
- Need $q = 1$ to control variance when sampling classes (next).

Entrywise ℓ_1-norm: $\|U\|_{\mathscr{U}} = \sum_{j\in[d]} \sum_{\kappa\in[k]} |U_{j\kappa}|.$

- Convert $\mathcal{U} = \{U : \|U\|_1 \leqslant R^*\}$ into *"solid"* simplex in $\mathbb{R}^{2d\times k}$:

$$\begin{aligned} \widehat{\mathcal{U}} = & \quad \{\widehat{U} \in \mathbb{R}^{2d\times k} : \mathbb{1}_{2d\times k}^\top \widehat{U} \leq R^*\}, \\ \Leftrightarrow & \quad \widehat{U} = [U_+; U_-]. \end{aligned}$$



Potential: renormalized nonnegative entropy (1-strongly cvx on $\widehat{\mathcal{U}}$),

Partial and full sampling

$$G(U^t, V^t) = \frac{1}{n}[{}_d X^\top {}_n (V^t - Y)_k, -{}_n X_d U^t_k].$$

Computing $G(U^t, V^t)$ is **hard:** $O(dnk)$ arithmetic operations (a.o.).

- Sample the *rows* of U^t and $V^t - Y$, i.e. **features** and **examples**:

$$\xi_U(\mathsf{p}) = X \frac{e_j e_j^\top}{\mathsf{p}_j} U, \quad \text{where } j \sim \mathsf{p} \in \Delta_d,$$

and similarly for $\eta_{V,Y}(\mathsf{q})$ with $\mathsf{q} \in \Delta_n$. Computed in $O(dk + nk)$.

Partial and full sampling

$$G(U^t, V^t) = \tfrac{1}{n}[{}_dX^\top{}_n(V^t - Y)_k, -{}_nX_d U^t_k].$$

Computing $G(U^t, V^t)$ is **hard:** $O(dnk)$ arithmetic operations (a.o.).

- Sample the *rows* of U^t and $V^t - Y$, i.e. **features** and **examples**:

$$\xi_U(\mathsf{p}) = X\frac{e_j e_j^\top}{\mathsf{p}_j}U, \quad \text{where } j \sim \mathsf{p} \in \Delta_d,$$

  and similarly for $\eta_{V,Y}(\mathsf{q})$ with $\mathsf{q} \in \Delta_n$. Computed in $O(dk + nk)$.
- Further sample **classes** (conditionally):

$$\xi_U(\mathsf{p}, \mathsf{P}) = X\frac{e_j e_j^\top}{\mathsf{p}_i}U\frac{e_\kappa e_\kappa^\top}{\mathsf{P}_{j\kappa}}, \quad \text{where } \begin{cases} j \sim \mathsf{p}, \\ \kappa \sim \mathsf{P} \in \Delta_k^{\otimes d}, \end{cases}$$

  and similarly for $\eta_{V,Y}(\mathsf{q}, \mathsf{Q})$ with $\mathsf{Q} \in \Delta_k^{\otimes n}$.
- Computed in $O(d + n + k)$ a.o. as we only need one row of P, Q.

Variance control

Ideally, we are interested in solving

$$\inf_{\mathsf{p}\in\Delta_d,\mathsf{P}\in\Delta_k^{\otimes d}} \left\{ \sup_{U\in\mathcal{U}} \mathbf{E}[\|XU - \xi_U(\mathsf{p})\|_{\mathscr{Y}*}^2] \right\}, \qquad \inf_{\mathsf{q}\in\Delta_n,\mathsf{Q}\in\Delta_k^{\otimes n}} \{\ldots\}.$$

Lemma. Second moments $\sup_{U\in\mathcal{U}} \mathbf{E}[\|\xi_U(\mathsf{p},\mathsf{P})\|_{\mathscr{Y}*}^2]$, ... minimized by:

$$\mathsf{p}_j^* \propto \|X(:,j)\|_2 \cdot \|U(j,:)\|_1, \qquad \mathsf{P}_{j\kappa}^* \propto |U_{j\kappa}|,$$
$$\mathsf{q}_i^* \propto \|X(i,:)\|_\infty \cdot \|V(i,:) - Y(i,:)\|_1, \qquad \mathsf{Q}_{i\kappa}^* \propto |V_{i\kappa} - Y_{i\kappa}|.$$

Final bound

With this choice of geometry and sampling distributions:

$$\mathbf{E}[\mathtt{Gap}^T] = \frac{R^*}{\sqrt{T}} \cdot \widetilde{O}\left(\max_{j\in[d]} \mathbf{E}_n^{1/2}[\phi_j^2] + \mathbf{E}_n\left[\max_{j\in[d]} |\phi_j|\right] \right).$$

For **light-tailed** data distribution, the two terms are of the same order.

Sublinear algorithm for multiclass SVM

Fully sampled oracle $\zeta_W = \frac{1}{n}[\eta_{V,Y}, -\xi_U]$ sparse, cost $O(d+n+k)$.

Challenges:

- Iterate updates $W^{t+1} = \mathrm{prox}_{W^t}(\zeta_{W^t})$ **dense**.
- Maintain distributions p, q that depend on the row norms:

$$\mathrm{p}_j^* \propto \|X(:,j)\|_2 \cdot \|U(j,:)\|_1, \quad \mathrm{q}_i^* \propto \|X(i,:)\|_\infty \cdot \|V(i,:) - Y(i,:)\|_1.$$

Key idea: Entropy-type potentials + affine composite terms = special **multiplicative** updates: each row **rescaled** (modulo one element).

- Hence our trick with "solid" simplex / renormalized entropy in U.
- Hinge loss has affine dual composite term $\mathcal{F}_Y(V) = 1 - \mathrm{tr}[V^\top Y]$.

Sublinear algorithm for multiclass SVM, cont.

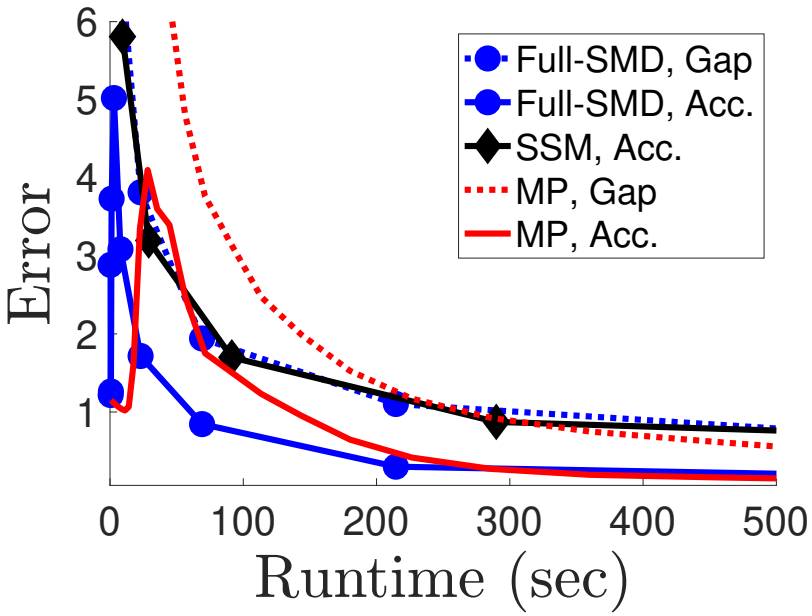
Algorithm sketch:

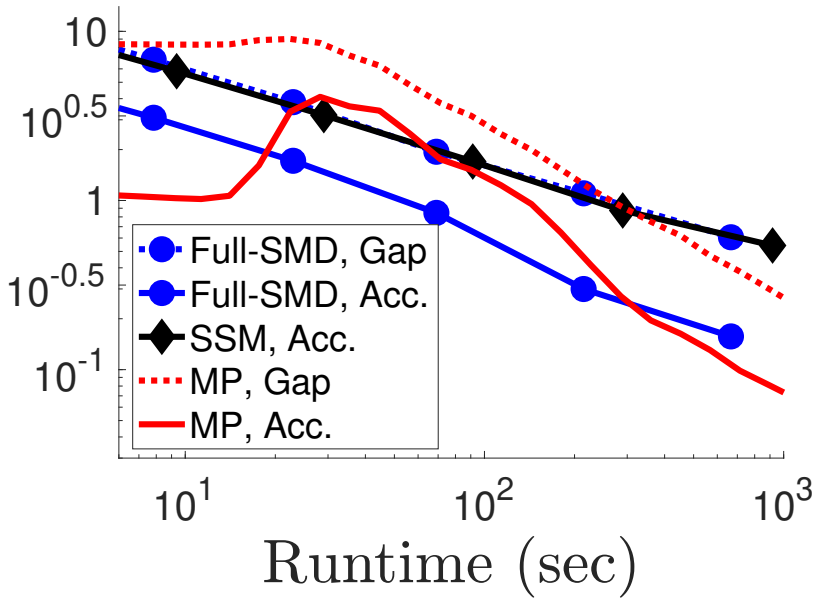
- Initialize $U^0 \in \Delta_{2dk}$, $V^0 \in \Delta_k^{\otimes n}$ by uniform distributions.
- Compute initial sampling distributions $\mathrm{p}^{(0)}, \mathrm{q}^{(0)}$ from X.
- For $t = 0, 1, ..., T$
  1. Sample $i_t \sim \mathrm{q}^{(t)}$, $j_t \sim \mathrm{p}^{(t)}$, $\kappa_t \sim \mathrm{Q}^{(t)}(i_t, :)$, $\kappa'_t \sim \mathrm{P}^{(t)}(j^t, :)$.
  2. Compute $\zeta_{W^t} = \frac{1}{n}[\eta_{V^t, Y}, -\xi_{U^t}]$ (two columns κ_t and κ'_t).
  3. **Lazy updates:** update scaling factors for the rows of U and V.
  4. Update $\|U(j_t, :)\|_1, \|V(i_t, :) - Y(i_t, :)\|_1$ and thus $\mathrm{p}^{(t)}$ and $\mathrm{q}^{(t)}$.
  5. Explicitly update the sampled element $U(j_t, \kappa_t)$, $V(i_t, \kappa'_t)$.

Behind the scenes: **lazy averaging** of iterates in $O(d + n + k)$ per iteration + $O(dk + nk)$ postprocessing.

- Sublinear iterations $O(d + n + k)$.
- Linear pre/postprocessing and memory: $O(dn + nk + dk)$.
- $O(dk)$ can be improved to $O(d \min(k, T))$.

Experiments and perspectives





Convergence of **sublinear stochastic mirror descent (Full-SMD)** vs. **mirror prox (MP)** and **SGD**, in natural (left) and logarithmic scale (right).

Perspectives:

(?) Faster (at least linear!) certificate for the duality gap.

(?) Logistic and other Fenchel-Young losses.

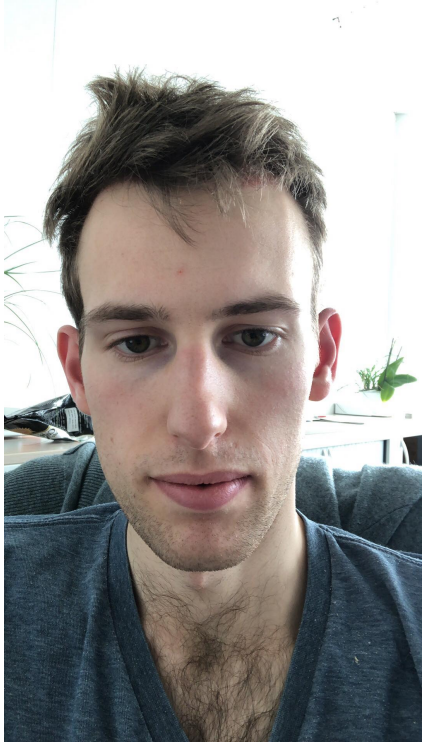
(?) Heavy-tailed data via Hadamard/DFT randomization technique.

Part II: Fast Rates for M-Estimators using Self-Concordance

D. M. Ostrovskii, F. Bach. Finite-Sample Analysis of M-Estimators using Self-Concordance. *arXiv:1810.06838*.

NOT IN THIS TALK:

U. Marteau-Ferey, D. M. Ostrovskii, A. Rudi, F. Bach.
Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance. *arXiv:1902.03046*.



Problem setup

Statistical learning problem

Given some **loss** $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, find a minimizer $\theta_* \in \mathbb{R}^d$ of **expected risk**:

$$\theta_* \in \operatorname*{Argmin}_{\theta \in \mathbb{R}^d} L(\theta) := \mathbf{E}[\ell(x^\top \theta, y)],$$

where expectation $\mathbf{E}[\cdot]$ is w.r.t. the unknown distribution $\mathcal{P}$ of $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$. Since $\mathcal{P}$ is unknown, θ_* can't be found; instead, it is estimated from **i.i.d. sample**:

$$(x_1, y_1), ..., (x_n, y_n) \sim \mathcal{P} \quad \text{(i.i.d.)}$$

- Random-design **classification**, $\mathcal{Y} = \{0, 1\}$, and **regression**, $\mathcal{Y} = \mathbb{R}$.
- Performance of a candidate $\widehat{\theta}$ measured by **excess risk** $L(\widehat{\theta}) - L(\theta_*)$.

Goal

- **Empirical risk minimization:** replace $L(\theta)$ with **empirical risk**:

$$\widehat{\theta}_n \in \operatorname*{Argmin}_{\theta \in \mathbb{R}^d} \left\{ L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^\top \theta, y_i) \right\}.$$

  Also called M-**estimation** in statistics.

- Special case: conditional **quasi maximum likelihood estimator** (qMLE):

$$\ell(\eta, y) = -\log p_\eta(y)$$

  for some parametric family $\{p_\eta(y), \eta \in \mathbb{R}\}$, possibly not including the true distribution $\mathcal{P}$

Goal

Extend the classical asymptotic theory to **finite-sample** setup.

Motivation 1: Classical asymptotic theory*

- **Local regularity assumptions**: $L(\theta)$ sufficiently smooth at θ_*, and

$$\mathbf{H} := \nabla^2 L(\theta_*) \succ 0.$$

- Gradient covariance $\mathbf{G} := \mathbf{E}[\nabla_\theta \ell(x^\top \theta_*, y)\, \nabla_\theta \ell(x^\top \theta_*, y)^\top]$, and let

$$\mathbf{M} := \mathbf{H}^{-1/2} \mathbf{G} \mathbf{H}^{-1/2}.$$

$d_{\text{eff}} := \text{tr}(\mathbf{M})$ is the **effective dimension**. In well-specified models:

$$\mathbf{G} = \mathbf{H} \;\Rightarrow\; \mathbf{M} = \mathbf{I}_d \;\Rightarrow\; d_{\text{eff}} = d.$$

- In the limit $n \to \infty$, Central Limit Theorem & Taylor Expansion give:

$$\sqrt{n}\mathbf{H}^{-1/2}(\widehat{\theta}_n - \theta_*) \rightsquigarrow \mathcal{N}(0, \mathbf{M}),$$
$$n\|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \rightsquigarrow \mathcal{N}(0, \mathbf{M})^2, \quad 2n(L(\widehat{\theta}_n) - L(\theta_*)) \rightsquigarrow \mathcal{N}(0, \mathbf{M})^2.$$

$$\left\{ L(\widehat{\theta}_n) - L(\theta_*),\, \|\mathbf{H}^{1/2}(\theta_n - \theta_*)\|^2 \right\} = O\left(\frac{d_{\text{eff}} \log(1/\delta)}{n} \right).$$

*[Borovkov, 1998; van der Vaart, 1998; Lehmann and Casella, 2006].

Motivation 2: Random-design linear regression, I

- Gaussian model $y = \mathcal{N}(x^\top \theta, \sigma^2)$ leads to quadratic loss and risk:

$$\ell(x^\top \theta, y) = \tfrac{1}{2\sigma^2}(y - x^\top \theta)^2,$$
$$L(\theta) - L(\theta_*) = \tfrac{1}{2}\|\mathbf{H}^{1/2}(\theta - \theta_*)\|^2,$$
$$L_n(\theta) - L_n(\theta_*) = \tfrac{1}{2}\|\mathbf{H}_n{}^{1/2}(\theta - \theta_*)\|^2 + \underbrace{\langle \nabla L_n(\theta_*), \theta - \theta_* \rangle}_{\text{zero-mean}}$$

- In particular, at any θ we have $\nabla^2 L(\theta) \equiv \mathbf{H}$ and $\nabla^2 L_n(\theta) \equiv \mathbf{H}_n$ with

$$\mathbf{H} = \mathbf{E}[xx^\top], \quad \mathbf{H}_n = \tfrac{1}{n}\sum_{i=1}^n x_i x_i^\top.$$

Motivation 2: Random-design linear regression, I

- Gaussian model $y = \mathcal{N}(x^\top \theta, \sigma^2)$ leads to quadratic loss and risk:

$$\ell(x^\top \theta, y) = \tfrac{1}{2\sigma^2}(y - x^\top \theta)^2,$$
$$L(\theta) - L(\theta_*) = \tfrac{1}{2}\|\mathbf{H}^{1/2}(\theta - \theta_*)\|^2,$$
$$L_n(\theta) - L_n(\theta_*) = \tfrac{1}{2}\|\mathbf{H}_n{}^{1/2}(\theta - \theta_*)\|^2 + \underbrace{\langle \nabla L_n(\theta_*), \theta - \theta_* \rangle}_{\text{zero-mean}}$$

- In particular, at any θ we have $\nabla^2 L(\theta) \equiv \mathbf{H}$ and $\nabla^2 L_n(\theta) \equiv \mathbf{H}_n$ with

$$\mathbf{H} = \mathbf{E}[xx^\top], \quad \mathbf{H}_n = \tfrac{1}{n}\sum_{i=1}^n x_i x_i^\top.$$

Theorem T: Estimation of a sample covariance matrix [Vershynin, 2010]

Assume $\mathbf{H}^{-1/2}X$ is subgaussian, i.e., has tails lighter than $\mathcal{N}(\mu, \mathbf{I}_d)$, and

$$n \gtrsim d + \log(1/\delta).$$

Then, with probability at least $1 - \delta$ it holds:

$$0.5\mathbf{H} \preccurlyeq \mathbf{H}_n \preccurlyeq 2\mathbf{H}.$$

Motivation 2: Random-design linear regression, II

Theorem 0: Finite-sample risk bound for linear regression [Hsu et al., 2012]

Assume that $\mathbf{H}^{-1/2}x$ and $\mathbf{G}^{-1/2}\nabla\ell_\theta(x^\top\theta_*, y)$ are subgaussian, and

$$n \gtrsim d + \log(1/\delta).$$

Then w.p. at least $\geq 1-\delta$,

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \|\mathbf{H}^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim \frac{d_{\text{eff}}\log(1/\delta)}{n}.$$

Motivation 2: Random-design linear regression, II

Theorem 0: Finite-sample risk bound for linear regression [Hsu et al., 2012]

Assume that $\mathbf{H}^{-1/2}x$ and $\mathbf{G}^{-1/2}\nabla\ell_\theta(x^\top\theta_*, y)$ are subgaussian, and

$$n \gtrsim d + \log(1/\delta).$$

Then w.p. at least $\geq 1-\delta$,

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \|\mathbf{H}^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim \frac{d_{\text{eff}}\log(1/\delta)}{n}.$$

Proof sketch:

1. Since $\nabla L_n(\widehat{\theta}_n) = 0$, we have $\|\mathbf{H}_n^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 = \|\mathbf{H}_n^{-1/2}\nabla L_n(\theta_*)\|^2$.
2. Combining with **Theorem T**,

$$L(\widehat{\theta}_n) - L(\theta_*) = \tfrac{1}{2}\|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \leq 2\|\mathbf{H}^{-1/2}\nabla L_n(\theta_*)\|^2;$$

3. The score $\mathbf{H}^{-1/2}\nabla L_n(\theta_*)$ is the average of n i.i.d. subgaussian vectors.

Towards the general case

- Generally, risk is not quadratic, and Hessians are not constant:

$$\nabla^2 L(\theta) = \mathbf{H}(\theta), \quad \nabla^2 L_n(\theta) = \mathbf{H}_n(\theta).$$

- To extend the previous argument, we must control the precision of **local quadratic approximation** of $L_n(\theta)$ and $L(\theta)$ around θ_*.
- We exploit **self-concordance**, a concept introduced in [Nesterov and Nemirovski, 1994] in the theory of interior-point methods, and brought to the statistical learning context in [Bach, 2010] to study logistic regression.

Self-concordant losses

We always assume that $\ell(\eta, y)$ is convex in η.

Definition. $\ell(\eta, y)$ is **self-concordant (SC)** if for any $(\eta, y) \in \mathbb{R} \times \mathcal{Y}$ it holds

$$|\ell'''_\eta(\eta, y)| \leq [\ell''_\eta(\eta, y)]^{3/2}.$$

Self-concordant losses

We always assume that $\ell(\eta, y)$ is convex in η.

Definition. $\ell(\eta, y)$ is **self-concordant (SC)** if for any $(\eta, y) \in \mathbb{R} \times \mathcal{Y}$ it holds

$$|\ell'''_\eta(\eta, y)| \le [\ell''_\eta(\eta, y)]^{3/2}.$$

- While the above definition is homogeneous in η, the next one is not:

Definition. $\ell(\eta, y)$ is **pseudo self-concordant (PSC)** if instead it holds

$$|\ell'''_\eta(\eta, y)| \le \ell''_\eta(\eta, y).$$

- **PSC** losses are somewhat more common than **SC** ones.
- However, obtaining optimal rate for **PSC** losses requires larger sample size.

Example 1: Generalized linear models

Conditional negative log-likelihood of y given $\eta = x^\top \theta$ in the form

$$\ell(\eta, y) = -y\eta + a(\eta) - b(y),$$

where $a(\eta)$ is called the **cumulant**, and is given by

$$a(\eta) = \log \int_{\mathcal{Y}} e^{y\eta + b(y)} \mathrm{d}y.$$

This defines the density $p_\eta(y) \propto e^{y\eta + b(y)}$ such that $a(\eta) = \mathbf{E}_{p_\eta}[y]$.

PSC: **Logistic regression** since $(\mathcal{Y} = \{0, 1\})$, and

$$|a'''(\eta)| = |\mathbf{E}_{p_\eta}(y - \mathbf{E}_{p_\eta}[y])^3| \leq \mathbf{E}_{p_\eta}[(y - \mathbf{E}_{p_\eta}[y])^2] = a''(\eta).$$

PSC: **Poisson regression:** $Y \sim \mathrm{Poisson}(e^\eta)$, then $a(\eta) = \exp(\eta)$.
SC: **Exponential-response model:** $Y \sim \mathrm{Exp}(\eta)$, $\eta > 0$, $a(\eta) = -\log(\eta)$.

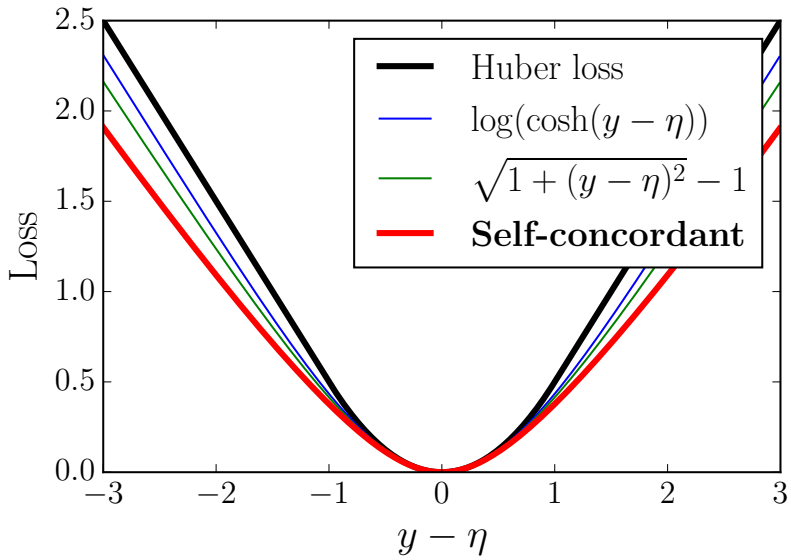
Example 2: Robust estimation

Loss $\ell(y, \eta) = \varphi(y - \eta)$ with $\varphi(t)$ convex, even, 1-Lipschitz, and $\varphi''(0) = 1$.

- **Huber loss**

$$\varphi(t) = \begin{cases} t^2/2, & |t| \le 1, \\ \tau t - 1/2, & |t| > 1. \end{cases}$$

$\varphi''(t)$ discontinuous at ± 1.



PSC: **Pseudo-Huber losses:** $\varphi(t) = \log\cosh(t)$, $\varphi(t) = \sqrt{1 + t^2} - 1$.

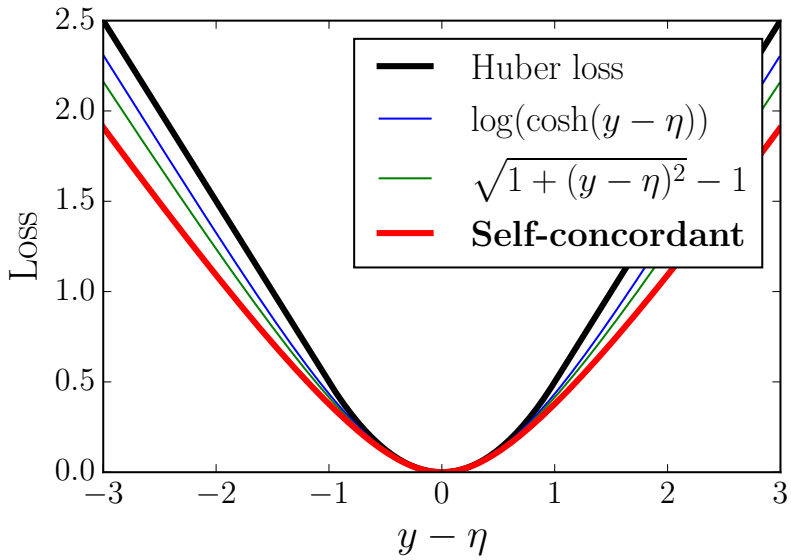
Example 2: Robust estimation

Loss $\ell(y, \eta) = \varphi(y - \eta)$ with $\varphi(t)$ convex, even, 1-Lipschitz, and $\varphi''(0) = 1$.

- **Huber loss**

$$\varphi(t) = \begin{cases} t^2/2, & |t| \le 1, \\ \tau t - 1/2, & |t| > 1. \end{cases}$$

$\varphi''(t)$ discontinuous at ± 1.



PSC: **Pseudo-Huber losses:** $\varphi(t) = \log\cosh(t)$, $\varphi(t) = \sqrt{1 + t^2} - 1$.

SC: **Fenchel dual of the log-barrier** $\phi(u) = -\log(1 - u^2)/2$ on $[-1, 1]$:

$$\varphi(t) = \frac{1}{2}\left[\sqrt{1 + 4t^2} - 1 + \log\left(\frac{\sqrt{1 + 4t^2} - 1}{2t^2}\right)\right].$$

Basic result

Recall that in the general case, we have the Hessian process $\mathbf{H}(\theta)$, given by

$$\mathbf{H}(\theta) := \mathbf{E}[\ell''(x^\top\theta, y)xx^\top] = \mathbf{E}[\widetilde{x}(\theta)\widetilde{x}(\theta)^\top],$$

where $\widetilde{x}(\theta) := [\ell''(x^\top\theta, y)]^{1/2}x$ is the *curvature-scaled design*.

Basic result

Recall that in the general case, we have the Hessian process $\mathbf{H}(\theta)$, given by

$$\mathbf{H}(\theta) := \mathbf{E}[\ell''(x^\top\theta, y)xx^\top] = \mathbf{E}[\widetilde{x}(\theta)\widetilde{x}(\theta)^\top],$$

where $\widetilde{x}(\theta) := [\ell''(x^\top\theta, y)]^{1/2}x$ is the *curvature-scaled design*.

Theorem 1: Finite-sample excess risk bound for self-concordant losses

Assume that the loss is **SC**, and $\mathbf{G}^{-1/2}\nabla\ell_\theta(x^\top\theta_*, y)$ and $\mathbf{H}(\theta_*)^{-1/2}\widetilde{x}(\theta_*)$ are subgaussian. Whenever

$$n \gtrsim d + \log(1/\delta) \vee d_{\text{eff}}\, d \log(1/\delta),$$

with probability $1-\delta$ it holds

$$L(\widehat{\theta}_n) - L(\theta_*) \lesssim \|\mathbf{H}^{1/2}(\widehat{\theta}_n - \theta_*)\|^2 \lesssim \|\mathbf{H}^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim \frac{d_{\text{eff}}\log(1/\delta)}{n}.$$

- ☺ Distribution conditions are local (only at θ_*);
- ☹ Large sample complexity – scaling as the product $O(d_{\text{eff}}\, d)$.

Analysis: Key observation

Given $\mathbf{H}(\theta) = \nabla^2 L(\theta)$, consider **Dikin ellipsoids** of $L(\theta)$ at θ_0:

$$\Theta(\theta_0, r) := \{\theta : \|\mathbf{H}(\theta_0)^{1/2}(\theta - \theta_0)\|^2 \le r^2\}.$$

Key Observation. Suppose that $\mathbf{H}_n(\theta) \asymp \mathbf{H}_n(\theta_*)$ w.h.p. for any $\theta \in \Theta(\theta_*, r)$. Then, $\widehat{\theta}_n \in \operatorname{Argmin} L_n(\theta)$ belongs to $\Theta(\theta_*, r)$ once

$$\|\mathbf{H}(\theta_*)^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim r^2,$$

Proof sketch:

- Indeed, by definition of $\widehat{\theta}_n$, $L_n(\widehat{\theta}_n) \le L_n(\theta_*)$. Assume $\widehat{\theta}_n \notin \Theta_n(\theta_*, r)$.
- Pick $\bar{\theta}_n \in [\theta_*, \widehat{\theta}_n]$ **on the border of** $\Theta_n(\theta_*, r)$. Still, $L_n(\bar{\theta}_n) \le L_n(\theta_*)$.

$$0 \ge L_n(\bar{\theta}_n) - L_n(\theta_*) \approx \langle \nabla L_n(\theta_*), \bar{\theta}_n - \theta_* \rangle + \underbrace{\|\mathbf{H}_n(\theta_*)^{1/2}(\bar{\theta}_n - \theta_*)\|^2}_{\approx r^2 \text{ (by Theorem T)}}.$$

- By Cauchy-Schwarz, we arrive at $\|\mathbf{H}(\theta_*)^{-1/2}\nabla L_n(\theta_*)\|^2 \gtrsim r^2$.

Contradiction! ■

Analysis: recap

- Once $\widehat{\theta}_n$ has been localized to the neighborhood of θ_* where $L_n(\theta)$ is quadratic, we can mimick the argument for linear regression.
- Localization is guaranteed once

$$\|\mathbf{H}(\theta_*)^{-1/2}\nabla L_n(\theta_*)\|^2 \lesssim r^2,$$

which leads to the second threshold for n:

$$n \gtrsim \frac{1}{r^2} d_{\text{eff}} \log(1/\delta).$$

- **Now the question is:**

What is the radius r of the Dikin ellipsoid in which $\mathbf{H}_n(\theta) \approx \mathbf{H}_n(\theta_*)$?

  - **Short answer:** we can afford $r^2 \approx 1/d$ using self-concordance.

Analysis: self-concordance at play

What is the radius r of the Dikin ellipsoid in which $\mathbf{H}_n(\theta) \approx \mathbf{H}_n(\theta_*)$?

1. Recall that
$$\mathbf{H}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ell''(x_i^\top \theta, y_i) x_i x_i.$$

2. Integrating $|\ell'''(\eta, y)| \le [\ell''(\eta, y)]^{\frac{3}{2}}$ from $\eta_* = x^\top \theta_*$ to $\eta = x^\top \theta$,
$$\frac{1}{(1 + [\ell''(\eta_*, y)]^{\frac{1}{2}} |\eta - \eta_*|)^2} \le \frac{\ell''(\eta, y)}{\ell''(\eta_*, y)} \le \frac{1}{(1 - [\ell''(\eta_*, y)]^{\frac{1}{2}} |\eta - \eta_*|)^2},$$
$$\frac{1}{(1 + |\langle \widetilde{x}(\theta_*), \theta - \theta_* \rangle|)^2} \le \frac{\ell''(x^\top \theta, y)}{\ell''(x^\top \theta_*, y)} \le \frac{1}{(1 - |\langle \widetilde{x}(\theta_*), \theta - \theta_* \rangle|)^2}.$$

3. The ratio is bounded when $|\langle \widetilde{x}(\theta_*), \theta - \theta_* \rangle| \le 1/2$, i.e., by Cauchy-Schwarz,
$$\underbrace{\|\mathbf{H}(\theta_*)^{-1/2} \widetilde{x}(\theta_*)\|}_{\approx \sqrt{d}} \cdot \underbrace{\|\mathbf{H}(\theta_*)^{1/2} (\theta - \theta_*)\|}_{r} \le 1/2 \;\Rightarrow\; \boxed{r \lesssim \frac{1}{\sqrt{d}}}. \;\blacksquare$$

Improved result

Theorem 2: Improved sample complexity for self-concordant losses

Assume the loss is **SC**, $\mathbf{G}^{-1/2}\nabla\ell_\theta(x^\top\theta_*, y)$ is subgaussian, and $\mathbf{H}(\theta)^{-1/2}\widetilde{x}(\theta)$ is subgaussian in the unit Dikin ellipsoid of $L(\theta)$ at θ_*:

$$\Theta(\theta_*, 1) = \{\theta : \|\mathbf{H}(\theta_*)^{1/2}(\theta - \theta_*)\| \le 1\}.$$

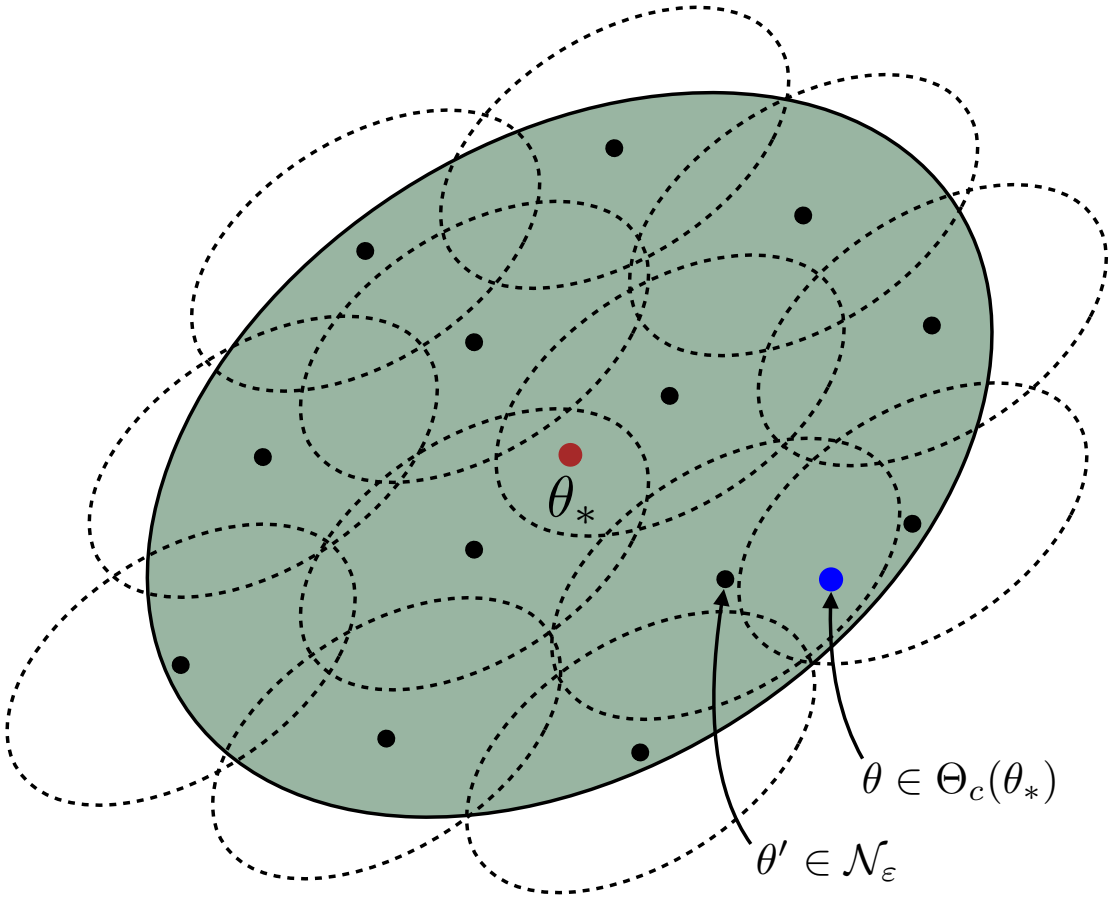
Then, the asymptotic bound holds already when

$$n \gtrsim d\log(d/\delta) \vee d_{\text{eff}}\log(1/\delta),$$

Main idea:

- Sample complexity $n \gtrsim d_{\text{eff}}\, d$ in Theorem 1 is due to Hessian approximation in the small Dikin ellipsoid with $r = O(1/\sqrt{d})$ rather than $r = O(1)$.
- We need to prove that $\mathbf{H}_n(\theta) \asymp \mathbf{H}_n(\theta_*)$ for $\theta \in \Theta(\theta_*, 1)$. To do this, we combine self-concordance with a **covering argument**.

Covering the Dikin ellipsoid



1. It is rather easy to prove first that $\mathbf{H}(\theta) \asymp \mathbf{H}(\theta_*)$ on $\Theta(\theta_*, 1)$.
2. By **SC**, $\mathbf{H}_n(\theta) \asymp \mathbf{H}_n(\theta_0)$ in a small ellipsoid $\Theta(\theta_0, 1/\sqrt{d})$.
3. Now cover $\Theta(\theta_*, 1)$ by $\Theta(\theta_0, 1/\sqrt{d})$ with θ_0 in the epsilon-net $\mathcal{N}_\varepsilon$. Control uniform deviations $\mathbf{H}_n(\theta)$ from $\mathbf{H}(\theta)$ on $\mathcal{N}_\varepsilon$. Note: $\log|\mathcal{N}_\varepsilon| = O(d \log d)$.

Pseudo self-concordant losses

- Because of the "incorrect" power of ℓ'' in **PSC**, we need an extra condition:

$$\mathbf{E}[xx^\top] \leq \rho \mathbf{E}[\ell''(x^\top \theta_*, y) xx^\top].$$

  for some $\rho > 0$. This condition is standard in logistic regression [Bach, 2010].
- We obtain similar results, but with ρ times worse sample complexity.
- Worst-case bounds on ρ can be exponentially bad [Hazan et al., 2014].
- However, this is not the case in practice. E.g., we show that

$$\rho \lesssim \|\theta_*\|_\Sigma^{3/2}$$

  in logistic regression with Gaussian design $x \sim \mathcal{N}(0, \Sigma)$.

Conclusion and perspectives

We use **self-concordance** – a concept from optimization – to obtain statistical results – **near-optimal** rates in finite-sample regimes in some statistical models.

Behind the scenes: regularized estimators (ℓ_2 and ℓ_1-regularization).

Perspectives:

- Heavy-tailed distributions
- Iterative algorithms: stochastic approximation, Quasi-Newton, ...
- Other models: covariance matrix estimation with log det loss, ...

Thank you!

References I

Bach, F. (2010). Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414.

Blondel, M., Martins, A. F. T., and Niculae, V. (2018). Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. *arXiv preprint arXiv:1805.09717*.

Borovkov, A. A. (1998). *Mathematical statistics*. Gordon and Breach Science Publishers.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654.

Grigoriadis, M. D. and Khachiyan, L. G. (1995). A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58.

Hazan, E., Koren, T., and Levy, K. Y. (2014). Logistic regression: tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209.

References II

Hsu, D., Kakade, S. M., and Zhang, T. (2012). Random design analysis of ridge regression. *The Journal of Machine Learning Research*, 23(9):1–24.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323.

Juditsky, A. and Nemirovski, A. (2011). First order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem's structure. *Optimization for Machine Learning*, pages 149–183.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2012). Block-coordinate frank-wolfe optimization for structural svms. *arXiv preprint arXiv:1207.4747*.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Nemirovski, A. and Yudin, D. (1983). Problem complexity and method efficiency in optimization.

Nesterov, Y. and Nemirovski, A. S. (1994). *Interior-point polynomial algorithms in convex programming*. Society of Industrial and Applied Mathematics.

References III

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*.

Xiao, L., Yu, A. W., Lin, Q., and Chen, W. (2017). DSCOVR: randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *arXiv preprint arXiv:1710.05080*.

Part III. Covariance Estimation for Heavy-Tailed Distributions

Joint work with **Alessandro Rudi**

D. M. Ostrovskii, A. Rudi. Affine-Invariant Covariance Estimation for Heavy-Tailed Distributions. *arXiv:1902.03086*.

Covariance Estimation Problem

Estimate the covariance matrix $\mathbf{S} = \mathbf{E}[XX^\top]$ from i.i.d. copies $X_1, ..., X_n$ of $X \in \mathbb{R}^d$.

- Sample covariance estimator:

$$\widetilde{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

- Relative spectral-norm guarantee: when X is light-tailed,

$$\frac{\|\widetilde{\mathbf{S}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \sqrt{\frac{\mathtt{r}(\mathbf{S})\log(d/\delta)}{n}} \quad \text{with probability} \quad \geq 1-\delta,$$

  where $\mathtt{r}(\mathbf{S}) = \frac{\mathrm{tr}(\mathbf{S})}{\|\mathbf{S}\|}$ is effective rank (Lounici & Kolchinskii 2014).
- Due to affine equivariance, this gives the guarantee

$$\left(1 - \sqrt{\frac{d\log(d/\delta)}{n}}\right)\mathbf{S} \preccurlyeq \widetilde{\mathbf{S}} \preccurlyeq \left(1 + \sqrt{\frac{d\log(d/\delta)}{n}}\right)\mathbf{S}.$$

Heavy-Tailed Distributions

$$\frac{\|\widetilde{\mathbf{S}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \sqrt{\frac{\mathtt{r}(\mathbf{S})\log(d/\delta)}{n}}$$

$$\left(1 - \sqrt{\frac{d\log(d/\delta)}{n}}\right)\mathbf{S} \preccurlyeq \widetilde{\mathbf{S}} \preccurlyeq \mathbf{S}\left(1 + \sqrt{\frac{d\log(d/\delta)}{n}}\right).$$

- The second guarantee is more useful in some applications (random-design linear regression, noisy PCA).
- Both require light-tailed assumptions on X, i.e. $\widetilde{\mathbf{S}}$ is not robust.
- Minsker (2014) proposes an estimator with a spectral-norm guarantee for heavy-tailed distributions (4th moment):

$$\widehat{\mathbf{S}}^{\mathrm{Min}} = \frac{1}{n}\sum_{i=1}^{n}\tau(\|X_i\|)X_iX_i^\top.$$

where $\tau(x)$ is the truncation map. Breaks affine equivariance!

Main Idea

- Minsker (2014) proposes an estimator with a spectral-norm guarantee for **heavy-tailed distributions** (4th moment):

$$\widehat{\mathbf{S}}^{\mathrm{Min}} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|X_i\|) X_i X_i^\top.$$

  where $\tau(x)$ is the truncation map.
- In fact, the desired $\preccurlyeq$ guarantee would hold for

$$\widehat{\mathbf{S}}^{*} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|\mathbf{S}^{-1/2} X_i\|) X_i X_i^\top,$$

  but it is unavailable, as $\mathbf{S}^{-1/2} X_i$'s are not observed.

Start with $\widehat{\mathbf{S}}_0 = \widehat{\mathbf{S}}^{\mathrm{Min}}$, and imitate $\widehat{\mathbf{S}}^{*}$ iteratively:

$$\widehat{\mathbf{S}}_{t+1} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \tau(\|\widehat{\mathbf{S}}_t^{-1/2} X_i\|) X_i X_i^\top,$$

Actual Estimator

$$\widehat{\mathbf{S}}_{t+1} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \tau(\|\widehat{\mathbf{S}}_t^{-1/2} X_i\|) X_i X_i^\top,$$

- Separate the sample $X_1, \ldots, X_n$ into batches, and use the new batch to compute $\widehat{\mathbf{S}}_{t+1}$.
- Iterative regularization: replace $\widehat{\mathbf{S}}_t^{-1/2}$ with $(\widehat{\mathbf{S}}_t + \lambda_t \mathbf{I})^{-1/2}$, where $\lambda_t = 2^{-t}\|\mathbf{S}\|$. Convergence in

$$O(\log(\mathrm{cond}(\mathbf{S}))) \text{ iterations,}$$

  where $\mathrm{cond}(\mathbf{S})$ is the condition number of $\mathbf{S}$.
- Similar complexity as for the sample covariance estimator!

Applications

- **Noisy PCA:** convergence of the noisy power method depends on the eigenvalue ratios which are controlled by the $\asymp$ guarantee.
- **Random-design linear regression**: we achieve optimal convergence rates in the setting with heavy-tailed design.

Convergence: deterministic term

Recall that $\Omega_{\mathcal{U}} = \widetilde{O}(R^{*2})$ and $\Omega_{\mathcal{V}} = \widetilde{O}(n)$.

Lemma.

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \frac{1}{n} \max_{j \in [d]} \|X_j\|_2,$$

where $X_j \in \mathbb{R}^n$ is the histogram of the j-th feature.

Corollary. Deterministic mirror descent converges with the rate

$$\frac{\mathcal{L}_{\mathscr{U},\mathscr{V}} \sqrt{\Omega_{\mathcal{U}} \Omega_{\mathcal{V}}}}{\sqrt{T}} = \widetilde{O}\left(\frac{R^*}{\sqrt{T}} \max_{j \in [d]} \sqrt{\frac{\|X_j\|_2^2}{n}} \right) = \widetilde{O}\left(\frac{R^*}{\sqrt{T}} \max_{j \in [d]} \mathbf{E}_n^{1/2}[\phi_j^2] \right),$$

where $\mathbf{E}_n^{1/2}[\phi_j^2]$ is the empirical 2nd moment of the j-th feature.

- When $n \to \infty$, $\max_{j \in [d]} \mathbf{E}_n^{1/2}[\phi_j^2] \to \max_{j \in [d]} \mathbf{E}_*^{1/2}[\phi_j^2]$ by LLN.
- We have equivalence in finite-sample when data is **light-tailed**.

Variance control (full)

Ideally, we are interested in solving

$$\inf_{\mathsf{p}\in\Delta_d,\mathsf{P}\in\Delta_k^{\otimes d}} \left\{ \sup_{U\in\mathcal{U}} \mathbf{E}[\|XU-\xi_U(\mathsf{p})\|_{\mathcal{Y}*}^2] \right\}, \qquad \inf_{\mathsf{q}\in\Delta_n,\mathsf{Q}\in\Delta_k^{\otimes n}} \{\dots\}.$$

Lemma. Second moments $\sup_{U\in\mathcal{U}} \mathbf{E}[\|\xi_U(\mathsf{p},\mathsf{P})\|_{\mathcal{Y}*}^2]$, ... minimized by:

$$\mathsf{p}_j^* \propto \|X(:,j)\|_2 \cdot \|U(j,:)\|_1, \qquad \mathsf{P}_{j\kappa}^* \propto |U_{j\kappa}|,$$
$$\mathsf{q}_i^* \propto \|X(i,:)\|_\infty \cdot \|V(i,:)-Y(i,:)\|_1, \qquad \mathsf{Q}_{i\kappa}^* \propto |V_{i\kappa}-Y_{i\kappa}|.$$

Moreover, the corresponding variance proxies satisfy:

$$\sigma_{\mathcal{U}}^2 \leqslant 4R^{*2}\mathcal{L}_{\mathcal{U},\mathcal{Y}}^2, \quad \sigma_{\mathcal{V}}^2 \leqslant 8n\mathcal{L}_{\mathcal{U},\mathcal{Y}}^2 + 8\mathbf{E}_n^2\left[\max_{j\in[d]}|\phi_j|\right]$$

Final bound

$$\mathbf{E}[\mathtt{Gap}^T] = \frac{R^*}{\sqrt{T}} \cdot \widetilde{O}\left(\max_{j\in[d]} \mathbf{E}_n^{1/2}[\phi_j^2] + \mathbf{E}_n\left[\max_{j\in[d]}|\phi_j|\right] \right).$$

For **light-tailed** data distribution, the two terms are of the same order.

Example 1: Generalized linear models (full)

Conditional negative log-likelihood of y given $\eta = x^\top \theta$ in the form

$$\ell(\eta, y) = -y\eta + a(\eta) - b(y),$$

where $a(\eta)$ is called the **cumulant**, and is given by

$$a(\eta) = \log \int_{\mathcal{Y}} e^{y\eta + b(y)} \mathrm{d}y.$$

This defines the density $p_\eta(y) \propto e^{y\eta + b(y)}$ such that $a(\eta) = \mathbf{E}_{p_\eta}[y]$, and

$$\ell_\eta^{(s)}(\eta, y) = a^{(s)}(\eta) = \mathbf{E}_{p_\eta}[(y - \mathbf{E}_{p_\eta} y)^s], \quad s \geq 2.$$

SC/**PSC** relate 2nd and 3rd central moments of $p_\eta(\cdot)$.

PSC: **Logistic regression** since $(\mathcal{Y} = \{0, 1\})$, and

$$|a'''(\eta)| = |\mathbf{E}_{p_\eta}(y - \mathbf{E}_{p_\eta}[y])^3| \leq \mathbf{E}_{p_\eta}[(y - \mathbf{E}_{p_\eta}[y])^2] = a''(\eta).$$

PSC: **Poisson regression:** $Y \sim \mathrm{Poisson}(e^\eta)$, then $a(\eta) = \exp(\eta)$.

SC: **Exponential-response model:** $Y \sim \mathrm{Exp}(\eta)$, $\eta > 0$, $a(\eta) = -\log(\eta)$.