Mirrorless Mirror Descent

Nati Srebro (TTIC)

Based on work with
Suriya Gunasekar (TTIC→MSR), Blake Woodworth (TTIC→ENS), Filip Hanzely (TTIC)

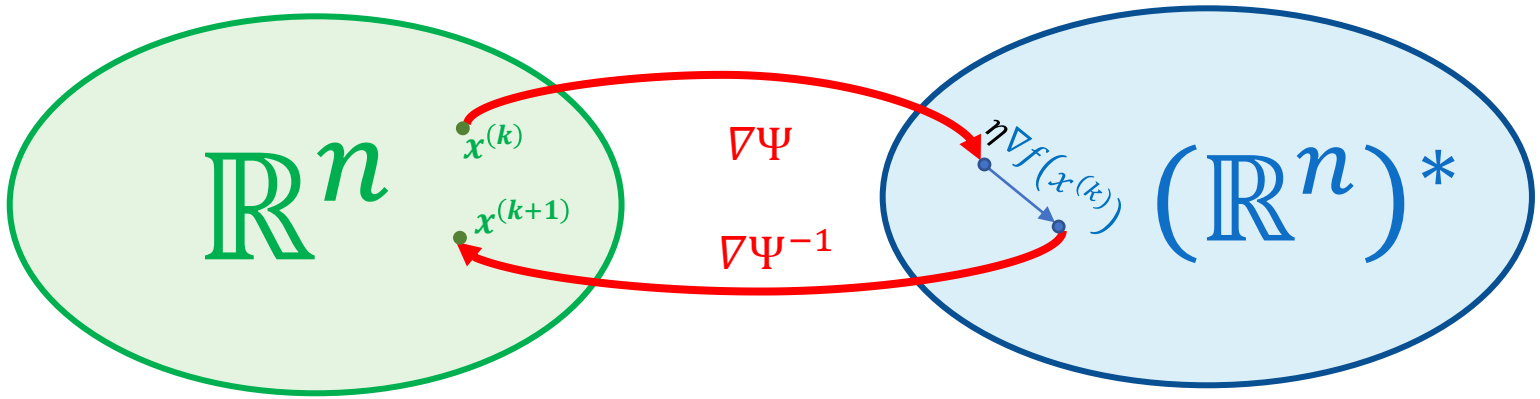
Plus: Bonus spotlight on acceleration, parallelization and interpolation learning!

A presentation in honor of Yurii Nesterov's 65th Birthday

$$\min_{x \in \mathbb{R}^d} f(x)$$

$$x_{k+1} \leftarrow \arg\min_x \langle \nabla f(x_k), x \rangle + \frac{1}{\eta_k} D_\Psi(x; x_k)$$
$$= \nabla\Psi^{-1}\left(\nabla\Psi(x_k) - \eta_k \nabla f(x_k)\right)$$

$$D_\Psi(x; y) = \Psi(x) - (\Psi(y) + \langle \nabla\Psi(y), x - y \rangle)$$



Classical analysis:

$$D_\Psi(x; y) \geq \tfrac{1}{2}\|x - y\|^2$$
$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|^2$$

Or, with relative smoothness:

$$\lambda\nabla^2\Psi \preccurlyeq \nabla^2 f \preccurlyeq L\nabla^2\Psi$$

[Bolte, Bauschke Teboulle][Lu Freund Nesterov]

➔

$$k \leq \frac{L D_\Psi(x_0; x^*)}{\epsilon_k}$$

$\epsilon_k = f(x_k) - f(x^*)$

$$k \leq \frac{L}{\lambda}\log\left(\frac{L D_\Psi(x_0; x^*)}{\epsilon_k}\right)$$

With restarts (non-montone stepsize):

[as in, e.g., Roulet d'Aspremont ‘17, Woodworth S ‘21]

$$k \leq \frac{L}{\lambda}\log\left(\frac{\epsilon_0}{\epsilon_k}\right)$$

$$\dot{x}(t) = -\ \boldsymbol{H}\ \left(x(\lfloor t \rfloor_\eta)\right)^{-1} \nabla f\left(x(\lfloor t \rfloor_\eta)\right)$$

$\lfloor t \rfloor_\eta = \eta \left\lfloor t/\eta \right\rfloor$

$$\boldsymbol{H}(x) = \nabla^2 \Psi(x)$$

$\eta \to 0$

Forward Euler
Full Discretization
$x_k = x(k\eta)$

$$x_{k+1} \leftarrow \nabla\Psi^{-1}\left(\nabla\Psi(x_k) - \eta\, \nabla f(x_k)\right)$$

$$x_{k+1} \leftarrow x_k - \eta\ \boldsymbol{H}\ (x_k)^{-1} \nabla f(x_k)$$

Forward Euler

$$\frac{d}{dt} \nabla\Psi\left(x(t)\right) = -\nabla f\left(x(t)\right)$$

"Natural Gradient Descent" [Amari]

$$\dot{x}(t) = -\boldsymbol{H}\left(x(t)\right)^{-1}\nabla f\left(x(\lfloor t \rfloor_\eta)\right)$$

$\lfloor t \rfloor_\eta = \eta \left\lfloor t/\eta \right\rfloor$

$$\boldsymbol{H}(x) = \nabla^2\Psi(x)$$

Partial Discretization
$x_k = x(k\eta)$

$$x_{k+1} \leftarrow \nabla\Psi^{-1}\left(\nabla\Psi(x_k) - \eta\,\nabla f(x_k)\right)$$

"Mirror Descent" / Bregman Gradient Scheme

Forward Euler
Full Discretization
$x_k = x(k\eta)$

$$x_{k+1} \leftarrow x_k - \eta\,\boldsymbol{H}\,(x_k)^{-1}\nabla f(x_k)$$

"Natural Gradient Descent" [Amari]

Proof: define $x(t) \doteq \nabla\Psi^{-1}\left(\nabla\Psi(x_k) - (t - k\eta)\,\nabla f(x_k)\right)$

$k\eta \le t \le (k+1)\eta$

$$\begin{aligned}\dot{x}(t) &= \nabla^2\Psi^{-1}(\cdots)\left(-\nabla f(x_k)\right)\\ &= -\nabla^2\Psi\left(x(t)\right)^{-1}\nabla f(x(\lfloor t \rfloor_\eta)\end{aligned}$$

Potential-Free (Truly Primal, Mirrorless) Gradient Scheme

$$\dot{x}(t) = -\boldsymbol{H}\big(x(t)\big)^{-1}\nabla f\Big(x\big(\lfloor t \rfloor_\eta\big)\Big)$$
$$x_k = x(\eta k)$$

If $H(x)$ is a Hessian map: $H(x) = \nabla^2\Psi(x)$:

$$\begin{aligned} prox_{\Psi/\eta}(g,y) &\stackrel{\text{def}}{=} \arg min_x\, \eta\langle g, x\rangle + D_\Psi(x;y) \\ &= \nabla\Psi^{-1}(\nabla\Psi(y) - \eta g) \\ &= flow_{(\nabla^2\Psi)/\eta}(g,y) \end{aligned}$$

$$\begin{aligned} x_{k+1} &\leftarrow prox_{\Psi/\eta}(\nabla f(x_k), x_k) \\ &= flow_{(\nabla^2\Psi)/\eta}(\nabla f(x_k), x_k) \end{aligned}$$

For any smooth $H(x) \succ 0$:

$flow_{H/\eta}(g,y) \stackrel{\text{def}}{=} x(\eta)$ where x is the solution to

$$\dot{x}(\tau) = -H\big(x(\tau)\big)^{-1} g \qquad x(0) = y$$

$$x_{k+1} \leftarrow flow_{H/\eta}(\nabla f(x_k), x_k)$$

Relative smoothness (and strong convexity):

$$\lambda H(x) \preccurlyeq \nabla^2 f(x) \preccurlyeq L H(x)$$

If H is a Hessian map: $k \le \frac{L}{\lambda}\log\left(\frac{\epsilon_0}{\epsilon_k}\right)$

????? $k \le \frac{L}{\lambda}\log\left(\frac{\epsilon_0}{\epsilon_k}\right)$?????

Change of parameterization:

$$\tilde{x} = \phi(x)$$

$$\tilde{f}(\tilde{x}) = f\left(\phi^{-1}(\tilde{x})\right)$$

$$\tilde{H}(\tilde{x}) = \nabla\phi^{-1}(\tilde{x}) H\left(\phi^{-1}(\tilde{x})\right) \nabla\phi^{-1}(\tilde{x})$$

$$\min_{x} f(x) \qquad\qquad \min_{\tilde{x}} \tilde{f}(\tilde{x})$$

$$(MD) \qquad \dot{x}(t) = -H\big(x(t)\big)^{-1} \nabla f\left(x\left(\lfloor t \rfloor_{\eta}\right)\right) \qquad\qquad \dot{\tilde{x}}(t) = -\tilde{H}\big(\tilde{x}(t)\big)^{-1} \nabla \tilde{f}\left(\tilde{x}\left(\lfloor t \rfloor_{\eta}\right)\right) \qquad (\widetilde{MD})$$

Is the solution invariant? Do the solutions of (MD) and $(\widetilde{MD})$ satisfy $\tilde{x}(t) = \phi\big(x(t)\big)$?
(It is invariant when $\eta = 0$)

Let $x(t)$ be the solution of (MD), set $\tilde{x}(t) \doteq \phi\big(x(t)\big)$ and check if $(\widetilde{MD})$ is satisfied:

$$\dot{\tilde{x}}(t) = -\tilde{H}\big(\tilde{x}(t)\big)^{-1} \left(\boldsymbol{\nabla\phi^{-1}}\left(\tilde{\boldsymbol{x}}\left(\lfloor \boldsymbol{t} \rfloor_{\eta}\right)\right) \boldsymbol{\nabla\phi}\left(\tilde{\boldsymbol{x}}(\boldsymbol{t})\right)\right) \nabla \tilde{f}\left(\tilde{x}\left(\lfloor t \rfloor_{\eta}\right)\right)$$

Mirrorless Mirror Descent

- Primal-Only, Potential-Free, derivation of Primal Gradient Scheme, as partial discretization of Riemannian Gradient Flow:

$$\dot{x}(t) = -H\big(x(t)\big)^{-1}\nabla f\left(x\left(\lfloor t \rfloor_{\eta}\right)\right)$$
$$x^{(k)} = x(\eta k)$$

- Valid first-order method for any metric tensor $H(x)$
- Does relative smoothness guarantee hold even if H is not a Hessian map?

$$\lambda H(x) \preccurlyeq \nabla^2 f(x) \preccurlyeq L H(x) \quad \rightarrow \quad k \leq \frac{L}{\lambda}\log\left(\frac{\epsilon_0}{\epsilon_k}\right) \quad ????$$

https://arxiv.org/abs/2004.01025

An Even More Optimal Stochastic Optimization Algorithm: Interpolation and Parallalization

Blake Woodworth (TTIC→ENS), Nati Srebro (TTIC)

$$\min F(x) \qquad F(x) = \mathbb{E}_{z\sim\mathcal{D}}[f(x,z)]$$

- Accelerated SGD: $f(x_k) - f(x^*) \le O\left(\frac{L\|x^*\|^2}{k^2} + \frac{\sigma\|x^*\|}{\sqrt{k}}\right)$ $\quad \forall_x \mathbb{E}[\|\nabla f(x,z_i) - \nabla F(x)\|^2] \le \sigma^2$ $\quad \nabla^2 f \preccurlyeq L$
- SGD (Without acceleration): $f(x_k) - f(x^*) \le O\left(\frac{L\|x^*\|^2}{k} + \frac{\sigma_*\|x^*\|}{\sqrt{k}}\right)$ $\quad \mathbb{E}[\|\nabla f(x^*,z_i) - \nabla F(x^*)\|^2] \le \sigma_*^2$
- Is it possible to get: $\frac{L\|x^*\|^2}{k^2} + \frac{\sigma_*\|x^*\|}{\sqrt{k}}$?

Interpolation (realizable) learning:
$f(x,z) \ge 0, F(x^*) = 0$ ➔ $\sigma_* = 0$

- Answer: **No!**
- For any stochastic optimization method (even if not first order) that uses k samples $z_1, \dots, z_k \sim iid\ \mathcal{D}$, there exists a problem with $f(x_k) - f(x^*) \ge \Omega\left(\frac{L\|x^*\|^2}{k} + \frac{\sigma_*\|x^*\|}{\sqrt{k}}\right)$
- But using k mini-batch grad estimates $g_i = \frac{1}{b}\sum_{j=1}^{b} \nabla f(x_i, z_{i,j})$,

  accelerated MB-SGD ensures: $f(x_k) - f(x^*) \le O\left(\frac{L\|x^*\|^2}{k^2} + \frac{L\|x^*\|^2}{bk} + \frac{\sigma_*\|x^*\|}{\sqrt{bk}}\right)$

  …and this is optimal (in terms of L, σ_*, b, for methods using mini-batch grad estimates)

Linear parallel speedup in broad regime (even when $\sigma_* = 0$)

When $\sigma_* = 0$,
parallelizing on $b = k$ machines yields optimal parallel runtime

An Even More Optimal Stochastic Optimization Algorithm, Woodworth Srebro 2021, https://arxiv.org/abs/2106.02720